White Paper

HOW DATA DEDUPLICATION WORKS AND WHY YOU NEED IT FOR YOUR SERVER BACKUPS

The Current Challenge

Enterprises are seeking new ways to keep up with the challenges of managing and protecting their rapidly expanding and distributed corporate data—especially data that resides on remote office servers as well as laptops, mobile devices (collectively known as endpoints). Currently, the amount of enterprise data is typically doubling every 14 months, the location of data is becoming more dispersed, and the linkage between data sets is becoming even more complex. These factors have a significant impact on storage and bandwidth costs.

Data deduplication helps enterprises dramatically reduce the amount of storage and bandwidth required for backups and other data-storage and data-access workloads. Druva uses a unique approach to data deduplication to help customers overcome these data challenges.

The Sharp Rise in Storage and Bandwidth Costs

Corporate data has increased sharply the past five years, which is driving a significant rise in bandwidth and storage costs. A survey by AFCOM (a data center trade organization) found that over 63% of IT managers surveyed have seen a dramatic increase in their storage costs. Two of the main reasons for this increase are the proliferation of mobile devices in enterprises and the more geographically dispersed nature of enterprises over the last decade. Fortunately—and unfortunately—for IT managers, much of the data increase is the direct result of replicated files across multiple infrastructure data repositories and endpoint devices.

> ***"Primary and secondary storage requirements are growing at around 40% year-over-year."***
>
> *—Jason Buffington, ESG Brief, "Plan for Hybrid Data Protection Media," January 2016*

Deduplication = Storage and Bandwidth Savings

Data deduplication is the elimination of redundant data. Deduplication algorithms identify and delete duplicate data, leaving only one copy (or a "single instance") of the data to be stored. However, indexing of all data is still retained should that data ever be needed. As a result, deduplication is able to reduce the required bandwidth and storage capacity.

> **Example:** A typical email system might contain 100 instances of the same 1MB file attachment. If the email platform is backed up or archived, all 100 instances are saved, requiring 100MB of storage space. With data deduplication, only one instance of the attachment is actually stored, and each subsequent instance is just referenced back to the one saved copy. In this example, a 100MB storage and bandwidth demand could be reduced to only 1MB.

The practical benefits of this technology depend on various factors, such as point of application, algorithm used, data type, and data retention and protection policies. Let's take a look at some of the ways deduplication technologies differ.

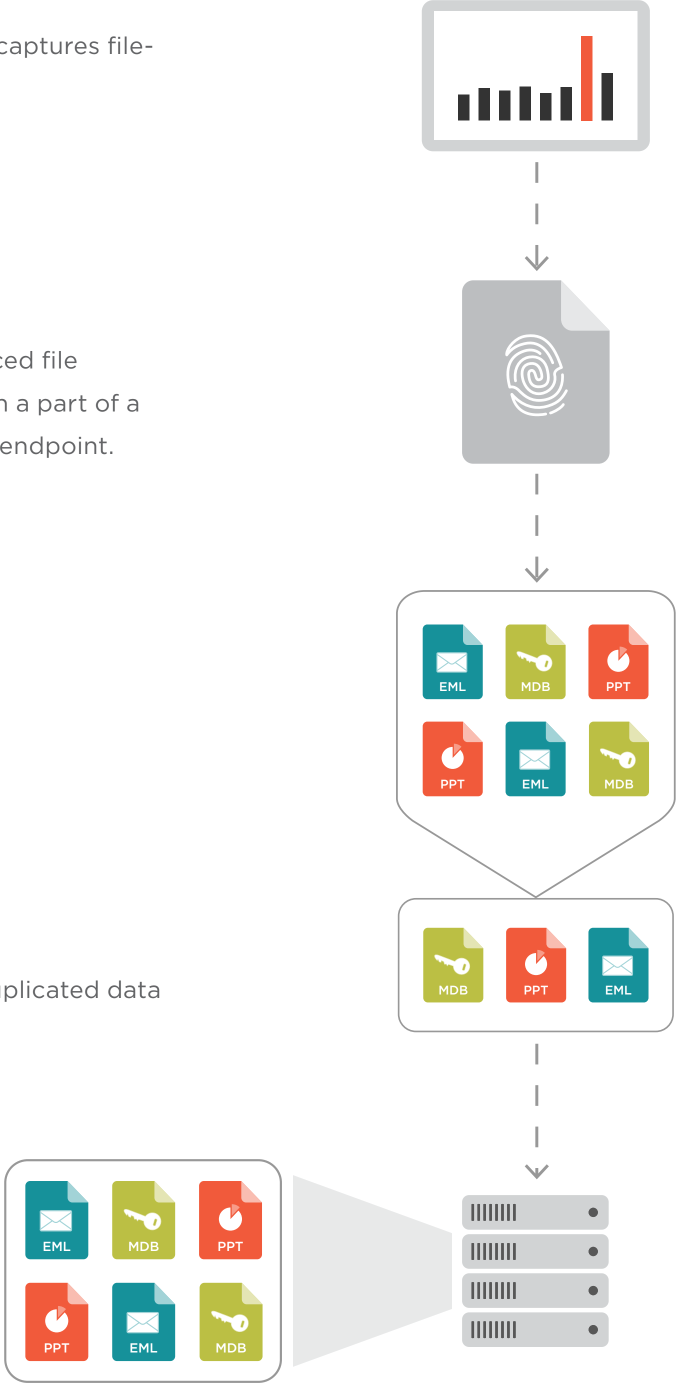
These technologies vary by:

- Where the deduplication happens (server or client-side)
- Granularity of the deduplication (file or subfile-based)
- The logic of discovering duplicate data (block-based or app-aware)

Druva's Industry-Leading Deduplication Technology

Druva's patented global-and app-aware deduplication technology provides unmatched bandwidth and storage savings for backup and file sharing.

Steps to Eliminate Duplicates in Backup



1 The Druva backup agent continuously monitors and captures file-level changes.

2 Before performing a backup, it uses patented advanced file fingerprinting to check with the server if a file or even a part of a file has been backed up before, by any server or any endpoint.

3 Then, it sends only the unique content to the server.

4 The backup server maintains only a single copy of duplicated data & multiple references.

5 During restore, a user sees all his files irrespective of the duplicates.

The benefits of this technology depend on the following factors:

- Point of application: source vs. target vs. global deduplication at source
- Time of application: inline vs. post-process
- Granularity: file vs. sub-file level
- Algorithm: fixed-size blocks vs. variable-length data segments

> *"Right now our global deduplication savings is 21.49 [to 1]."*
>
> *—Jessica Fletcher, IT Analyst Supervisor, Pall Corporation*

Global Deduplication

The client agent performs duplicate checks at the client device by comparing data from all enterprise users and from all their devices. Only a single instance of a block from across all devices/users is stored on the server. This combination of global and client-side deduplication provides an 80% savings in bandwidth and storage.

Client-Side Deduplication

Druva uses a client-triggered architecture with deduplication performed at the client—enabling high levels of scalability and security. By performing deduplication checks at the client, Druva is able to save substantial bandwidth. Additionally, client-side caching of these file-level and subfile-level deduplication checks makes backups much faster.

The client also has a powerful WAN optimization engine that can automatically prioritize network availability and set backup bandwidth as a percentage of the total available bandwidth. Druva ensures that a backup neither consumes a large percentage of the bandwidth nor disrupts the end-user experience.

Object-Based Application-Aware Deduplication

Druva understands the on-disk format of applications and uses this knowledge to significantly enhance the deduplication process, while guaranteeing 100% accuracy. The app-aware deduplication technology recognizes common applications such as Outlook (PST data files). App-aware deduplication eliminates the dependence on multiple checksums, resulting in faster deduplication. For other applications, Druva uses the variable-length, block-based deduplication methodology.

Deduplication within applications

Many applications tend to change the data structure of a file even if a small element of that data structure changes. As a result, the entire file appears different when stored persistently on disk.

Consider, for example, an Outlook PST file. The PST file changes up to 5% of the blocks, even if the user closes Outlook without any update. Druva's app-aware technology recognizes up to 87 different message types in PST to intercept the actual changes and back up only unique content (e.g., emails, attachments, calendar updates, etc.).

This approach **guarantees 100% deduplication accuracy** on supported applications and optimal use of storage and bandwidth.

Deduplication across applications

Each application stores data differently on a disk, and the representation of the data often changes completely once it is stored and indexed on the disk. A good example is an image file, which is present in a Word document as well as in a PST file as an attachment. Block-based deduplication is often unable to identify such duplicates across applications, because the data itself has changed.

Druva, on the other hand, understands the *logical view* of the data, so our app-aware technology is more efficient in discovering duplicates across applications and eradicating them than other deduplication based backup solutions.

Example: With app-aware deduplication, a Word document on a user's desktop can be easily identified as a duplicate of an email attachment that has been backed up and can be removed from the backup.

Unified Deduplication across Backup and File Sharing

Druva provides file sharing and collaboration as well as backup capabilities. The deduplication algorithm works across both of these data sets, further decreasing bandwidth and storage costs.

Salient Features of Druva's Deduplication Technology

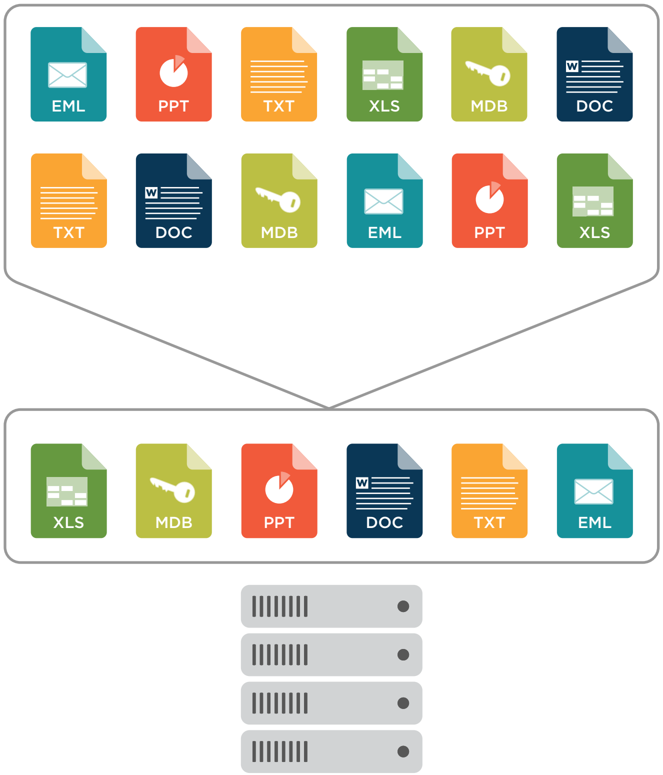
1 **Global:** data redundancy is eliminated across all users and endpoints.

2 **Client-side:** duplication checks and caching of these checks are performed at the client substantially reducing bandwidth and speeding up backups.

3 **Application Aware:** understanding of on-disk formats of applications, for example in Outlook, results in 100% accurate, faster deduplication and reduced storage requirements.

4 **High Performance:** Global and app-aware deduplication ensures only unique data is transferred to the server thereby resulting in up to 6x faster initial backups and optimal WAN bandwidth utilization.

5 **Unified Across Backup & File-Sharing:** deduplication works across backup and file-sharing functions to give massive storage & bandwidth savings.



A detailed explanation of deduplication techniques and Druva's innovative approach to deduplication is available in a white paper on data deduplication for endpoints at www.druva.com/resources.

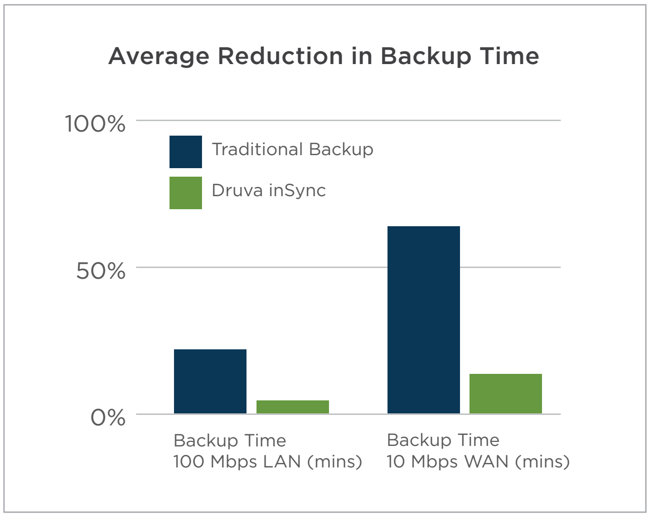
Organizations typically use two separate solutions for file sharing and backup. As a result, the same file may have multiple copies on the server, stored first by a backup utility and then again by a file-sharing tool. Even if the file-sharing tool in use might be capable of deduplication, it cannot perform deduplication across files stored by another tool, such as a backup utility. Druva is the only solution that allows an organization to use a single tool for both file sharing and backup. Because Druva is able to identify which files are stored for backup and which are for file sharing, it can eliminate duplications across the two functions and provide substantial bandwidth and storage cost advantages.

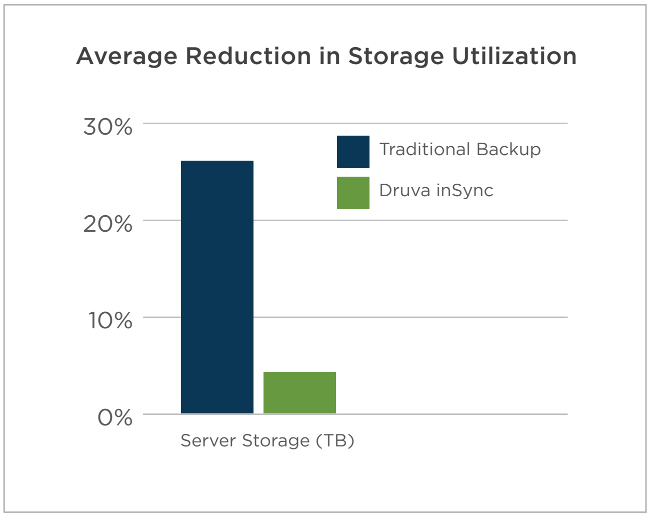
Bandwidth and Storage Savings from Druva's Deduplication

The following table benchmarks the performance of Druva's deduplication technology against four different customer installations in different industry verticals. These benchmarks clearly demonstrate the benefits delivered by Druva's global and app-aware deduplication technology in terms of backup time and storage utilization.

Example: If an email with a 1MB attachment is sent to 1,000 users, traditional incremental backup software would back up this 1MB attachment from each of the 1,000 different mailboxes. In contrast, Druva would back up 1MB from the first user and then skip all the other 999 copies as duplicates, saving over 99.9% backup time, bandwidth, and storage.

Customer	No of PCs	AVG. Backup Time LAN (min)		AVG. Backup Time VPN/WAN (min)		Total Storage Used (TB)	
		Old App	inSync	Old App	inSync	Old App	inSync
Large Financial Corp.	2,000	24	8	90	20	60	12
Oil & Gas Company	500	10	4	N/A	9	10	1.2
Consult. Group	300	15	6	40	8	27	2
Graphic Company	100	45	14	N/A	6	6.8	1.6





Druva vs. Other Deduplication Methods

Server-Side

The server-side deduplication method acts on the data on the server. In this case, the client is unaffected and does not benefit from any deduplication. The deduplication engine can be embedded in the hardware array, which can be used as a NAS/SAN device with deduplication capabilities. Alternatively, it can also be offered as an independent software or hardware appliance, which acts as an intermediary between the backup server and storage arrays.

However, in either case, this method does not decrease the amount of data transmitted or improve bandwidth utilization. It only improves storage utilization.

Client-Side

The client-side deduplication method acts on the data at the client (i.e., before it is moved to the server). A deduplication-aware backup agent is installed on the client, and the agent backs up only unique data. This approach results in improved bandwidth and storage utilization. However, this method imposes additional computational load on the backup client.

Druva’s Global Deduplication at the Source

No risk of data being lost with good restore speeds resulting from the significant bandwidth savings that come from not transmitting redundant data.

Fixed Block vs. Variable Block

Block-Based Deduplication

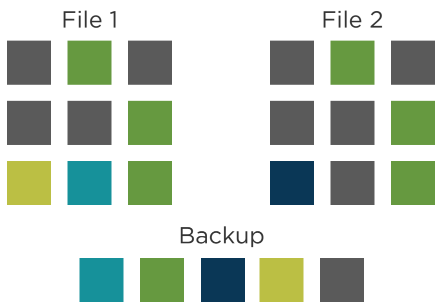
The block-based deduplication algorithms work as follows:

- The deduplication engine looks at a sequence of data, segments it into variable length blocks, and seeks blocks that are repeated.
- The engine stores a pointer to the original block instead of storing the duplicate block again.

There are two different approaches to block-based deduplication:

- Fixed-length block
- Variable-length data segment

Block-Based Deduplication



Fixed-Length Block

This approach divides the files into fixed-length blocks and uses a simple checksum-based method to find duplicates (e.g., MD5, SHA, etc.). Although it’s possible to look for repeated blocks, this approach provides very limited effectiveness since the primary opportunity for data reduction is in finding duplicate blocks in two transmitted datasets that are made up mostly—but not completely—of the same data segments.

Example: Similar data blocks may be present at different sets in two different datasets. In other words, the block boundary of similar data may be different.

This is very common when some bytes are inserted in a file, and when the changed file processes again and divides into fixed-length blocks. All blocks appear to have changed. Therefore, two datasets with a small amount of difference are likely to have very few identical fixed-length blocks.

Variable-Length Data Segment

This technology divides the data stream into variable-length data segments using a methodology that can find the same block boundaries in different locations and contexts. This allows the boundaries to “float” within the data stream so that changes in one part of the dataset have little or no impact on the boundaries in other locations of the dataset.

Through this method, duplicate data segments can be found at different locations within a file or between different files created by the same or a different application.

Limitations of Block-Based Deduplication

1 The block size (fixed or floating) used to determine data boundary is usually a “best” guess, & hence may not completely coincide with application’s actual block size.

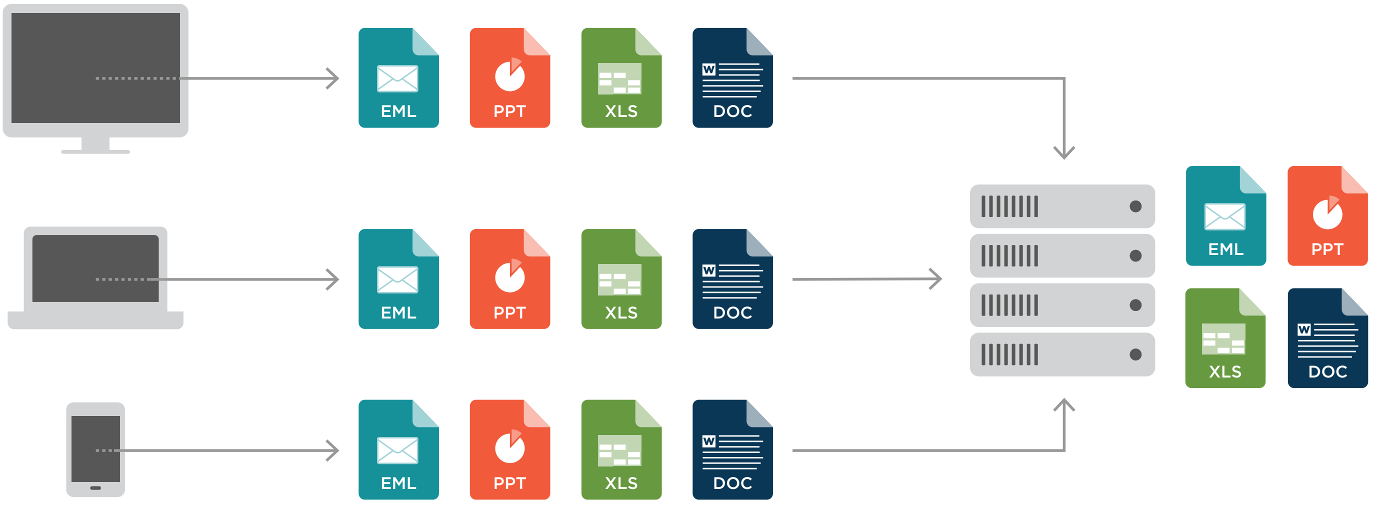
2 Different application have different ways of writing on-disk data, block-based algorithm will often fail to detect identical blocks across diffrent application file types (e.g. the same block of text stored in MS Word file and in a .PST email file).

3 Applications like Microsoft Outlook and Office use a complex database based on disk data structure which “stamp” each block with a unique header and footer, further complicating the task of finding duplicate blocks of data.

Druva’s Application-Aware Data Deduplication

Application-aware (app-aware) deduplication is a revolutionary concept that overcomes the limitations imposed by block-level deduplication. It benefits from knowledge of the format of data being backed up. Instead of guessing the optimal block size, it interprets the file as the actual application would, and identifies the logical blocks or messages within the files that have changed. As a result of understanding the structure of on-disk data, the deduplication algorithm is able to remove duplicates at a logical block or message level and is highly accurate. App-aware deduplication also brings performance improvements in the speed of deduplication because there is little or no scanning of data to find the “floating” block boundaries.

App-Aware Deduplication



Advantages of App-Aware Over Block-Based Deduplication

HIGHLY EFFICIENT

in removing duplicates across applications.

UP TO 300% MORE EFFICIENT

than simpler block-based approach in removing duplicates within complex applications like Microsoft Outlook.

UP TO 200% FASTER

data processing compared to the variable block-based approach.

This approach is highly efficient when it comes to complex applications like Microsoft Outlook and Office, which contribute to over 95% of the data on corporate PCs. App-aware deduplication not only identifies and removes all duplicates across emails and attachments within a single PST file, but it also effectively identifies and removes duplicates across different applications. Using this new approach, an image embedded within a Microsoft Word document can be identified as the duplicate of an image present as an attachment within a PST file.

Example: Assume a short email message is stored the following way on the disk.

```
Date 07.01.2014 ; From Bill Gates ; To: Warren Buffett ; Subject: Sell ; Body: sell everything!
```

Now assume that when Bill opens Outlook in a week time, and Outlook changes the date to 07/17 storing the message as:

```
Date 07.17.2014 ; From Bill Gates ; To: Warren Buffett ; Subject: Sell ; Body: sell everything!
```

Block-based deduplication methods will store the message in the following blocks before and after the change

BEFORE

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	BLOCK 5
Date 07.01.2014 ; Fro	m: Bill Gates ; To: W	arren Buffett : Sub	ject: Sell ; Body: s	ell everything!

AFTER

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	BLOCK 5
Date 07.17.2014 ; Fro	m: Bill Gates ; To: W	arren Buffett : Sub	ject: Sell ; Body: s	ell everything!

Note that changing the day designation in Block 1 from a single character (1) to two characters (17) causes all of the subsequent data to shift. As a result, the block-based deduplication engine will store a new copy of the message even though only one character was changed. Variable-block based deduplication does incrementally better in this case; however, it is still not optimal.

Hardware Deduplication vs. Cloud-Enabled Software Deduplication

Many organizations use virtual tape appliances or NAS devices for their deduplication needs because the appliances allow them to easily move from tape-based backup to disk-based backup. As data is written onto the appliance, data deduplication will normally occur in real-time. Since the compression process is handled by dedicated hardware rather than on a production server, virtual tape appliances can have considerably high data-compression ratios. Furthermore, because data stored on the appliance has already been deduplicated, large amounts of data can be sent across a WAN link in a short amount of time. Most modern backup appliances can write data directly to tape as well as copy and store data offsite or in the cloud.

While there are advantages to using hardware deduplication in certain cases, it also has some disadvantages. Hardware-based appliances are usually very expensive, and since deduplication is performed by physical appliances, architectural changes may be needed to accommodate the new hardware.

Cloud-Enabled Software Deduplication

Backup software may either perform source deduplication, which compresses data on the source server prior to performing a backup, or target deduplication, which compresses data on the backup server rather than on the source server. If deduplication is done at the source, less data needs to be sent across the network in the backup process, because the data is deduplicated before it is sent. With software deduplication, organizations can boost their backup speeds, which helps to lower storage and bandwidth costs. Furthermore, when backing up to the cloud, software deduplication reduces the amount of data that gets uploaded to the cloud.

Why go with cloud-native deduplication?

- No on-premises infrastructure
- Reduced costs for multiple sites
- Less IT overhead

Deduplication Designed for the Public Cloud

Druva's patented global deduplication happens at the source and at the block level, which enables endpoints and servers to benefit from efficient bandwidth utilization, minimizes the amount of data that needs to be transferred, and significantly improves RPOs and RTOs. A native cloud approach maximizes cloud elasticity and availability for any customer workload. It also avoids the bottlenecks of dedupe appliances and the restricted bandwidths of a single point of entry characterized by cloud-based, but not cloud-native, solutions.

Start exploring the benefits of the solution by visiting our Scale-Out Global Deduplication page.

About Druva

Druva is the global leader in Cloud Data Protection and Management, delivering the industry's first data management-as-a-service solution that aggregates data from endpoints, servers and cloud applications and leverages the public cloud to offer a single pane of glass to enable data protection, governance and intelligence—dramatically increasing the availability and visibility of business critical information, while reducing the risk, cost and complexity of managing and protecting it.

Druva's award-winning solutions intelligently collect data, and unify backup, disaster recovery, archival and governance capabilities onto a single, optimized data set. As the industry's fastest growing data protection provider, Druva is trusted by over 4,000 global organizations, and protects over 40 petabytes of data. Learn more at www.druva.com and join the conversation at twitter.com/druvainc.

druva

Druva, Inc.
Americas: +1 888-248-4976
Europe: +44 (0) 203-7509440
APJ: +919886120215
sales@druva.com
www.druva.com